Modeling Heterogeneity in Nest Survival Data

Ranjini Natarajan

Center for Statistical Sciences, 167 Angell Street, Box G-H,

Brown University, Providence, RI 02912, U.S.A.

and

Charles E. McCulloch

Biometrics Unit, Department of Statistical Science, 434 Warren Hall,

Cornell University, Ithaca, NY 14853, U.S.A.

SUMMARY

Current statistical methods for estimating nest survival rates assume that nests are identical in their propensity to succeed. However, there are several biological reasons to question this assumption. For example, experience of the nest builder, number of nest helpers, genetic fitness of individuals, and site-effects may contribute to an inherent disparity between nests with respect to their daily mortality rates (Klett and Johnson, 1982, *The Auk* **99**, 77-87). Ignoring such heterogeneity can lead to incorrect survival estimates. Our results show that constant survival models can seriously underestimate overall survival in the presence of heterogeneity. This paper presents a flexible random-effects approach to model heterogeneous nest survival data. We illustrate our methods through data on Red-winged Blackbirds and simulations.

Key words: Daily survival rate; Heterogeneity; Likelihood ratio test; Mayfield's method; Nest success; Numerical Integration; Over-dispersion; Random-effects.

1 Introduction

Statistical models for analyzing nest data commonly assume a constant probability of survival across nests (Mayfield, 1961; Bart and Robson, 1982; Pollock and Cornelius, 1988; Bromaghin and McDonald, 1993). However, there are several biological reasons to believe that daily mortality rates differ among nests, and not necessarily in relation to any variable the investigator is able to measure (Johnson, 1979; Klett and Johnson, 1982). For instance, the ability of nests to succeed may vary due to non-identifiable sources of heterogeneity, such as genetic fitness of individuals, or experience of the nest builder. Identifiable sources of variation can arise from differences in nest site, geographical location, etc.

Current approaches to account for variability in nest data are limited to factors which impact survival for all nests in a similar fashion. Pollock and Cornelius (1988) develop models where daily survival varies with the age of a nest, but is assumed to be constant across nests of the same age. Three limitations of this method are: (i) it cannot easily incorporate multiple factors (since sample sizes would become small with increasing stratification); (ii) it cannot include nest-level covariates; and (iii) the number of parameters can increase drastically with large nesting periods (as in waterfowl). A different approach to incorporate heterogeneity in survival has been proposed by Burnham and Rexstad (1993) using ultrastructure models superimposed on traditional models for band-recovery data.

This paper describes a flexible random-effects modeling approach to analyze nest survival data in the presence of tangible and intangible variation between nests. Our methods are applicable for estimating survival in any stage of the nest history, that is, incubation, or nestling. We first present a simple formulation to model "pure" heterogeneity, or non-identifiable sources of variation. This model can serve as a useful and quick diagnostic tool to evaluate the adequacy of a constant survival assumption. Next, we present a general random-effects model to estimate overall survival in the presence of covariates and multiple sources of heterogeneity. All notation for these models are initially presented assuming nests are found at the beginning of the stage. Extensions to the more

typical setting of encounter sampling in wildlife studies will also be discussed using the method of Bromaghin and McDonald (1993). Some advantages of our random-effects approach are it

1. allows calculation of predicted probabilities of survival for individual nests;

2. allows accommodation of a rich class of nest-level and time-varying covariate patterns, which facilitates a natural building of models for comparison (for example, comparing average survival for nests in two different habitats).

This paper is organized as follows: in Section 2 we describe a beta-binomial type model to capture pure heterogeneity in survival rates. A likelihood ratio statistic to test for inhomogeneities between nests is presented. In Section 3 we describe a random-effects model to account for multiple sources of heterogeneity and arbitrary structure of nest-level covariates. These methods are illustrated through data on Red-winged Blackbirds and simulations. Our results suggest that constant survival models can seriously underestimate overall survival in the presence of heterogeneity. Extensions to encounter sampling are discussed in Section 4.

2 A Simple Model for Pure Heterogeneity

Ornithologists visit nests periodically and monitor their survival status. Once found, nests are visited either until they fail or succeed (that is, hatch or fledge depending on the stage under study). The observed data for nest i is the pair $\{y_i, \ell_i\}$, where ℓ_i is the number of time units the nest is under observation and active, and y_i is a binary indicator of nest survival (coded as 1=survival; 0=failure). For the moment, we focus on the situation where all nests are found immediately after initiation, that is, ℓ denotes the length of life of a nest. (Throughout the manuscript we assume the absence of confounding temporal factors which allows us to re-align the cohort to all have the same time zero of initiation.) Then the contribution of nest i to the observed-data likelihood is:

$$f(y_i, \ell_i \mid p_i) = \begin{cases} p_i^{\ell_i} & \text{if } y_i = 1, \\ p_i^{\ell_i}(1-p_i) & \text{if } y_i = 0, \end{cases} \tag{1}$$

where $f(.)$ denotes probability density or mass functions, and p_i is the nest-specific survival probability for a single time-unit. The above formulation can be recognized as Mayfield's (1961) natural density, with the constant probability of success p replaced by nest-level survival rates p_i. The specification in (1) is completed by postulating a distribution for the p_i. A flexible and convenient assumption for the p_i is the beta distribution

$$f(p_i \mid \alpha, \beta) = \frac{1}{\mathrm{B}(\alpha, \beta)} p_i^{\alpha-1} (1-p_i)^{\beta-1}, \quad \alpha > 0, \ \beta > 0, \tag{2}$$

with: $\mathrm{B}(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$, mean $\mu_p = \alpha/(\alpha+\beta)$ and variance $v_p = \mu_p(1-\mu_p)/(\alpha+\beta+1)$ (Griffiths, 1973; Williams, 1975). An important modeling consequence of the specifications in (1) and (2) is the (non-negative) correlation induced between the survival status of a nest, at the repeated visits. Kahn and Raftery (1995) describe a similar formulation for studying hospital variation in the discharge of hip-fracture patients to skilled nursing facilities.

The estimation problem lies in calculating the probability of surviving a period of J (say) time-units, namely, $\mathcal{P} = E\left[p_i^J\right]$ where the expectation is over the survival distribution p_i. Upon noting that the marginal probability of surviving time-unit j ($j \geq 1$) given existence at the start of j is $(\alpha + j - 1)/(\alpha + \beta + j - 1)$, we have

$$\mathcal{P} = \prod_{j=1}^{J} \frac{(\alpha + j - 1)}{(\alpha + \beta + j - 1)}. \tag{3}$$

An estimator of survival and its' asymptotic variance (obtained by the delta method) under the heterogeneous model are:

$$\widehat{\mathcal{P}} = \prod_{j=1}^{J} \frac{(\widehat{\alpha} + j - 1)}{\left(\widehat{\alpha} + \widehat{\beta} + j - 1\right)},$$

$$\widehat{v(\widehat{\mathcal{P}})} = \underline{\widehat{\mathbf{d}}}^T(\widehat{\alpha}, \widehat{\beta})\, \underline{\widehat{\mathbf{V}}}(\widehat{\alpha}, \widehat{\beta})\, \underline{\widehat{\mathbf{d}}}(\widehat{\alpha}, \widehat{\beta}),$$

where $\widehat{\alpha}$, $\widehat{\beta}$ are the maximum likelihood estimates of α and β, $\underline{\widehat{\mathbf{V}}}$ is the associated variance matrix and $\underline{\widehat{\mathbf{d}}}$ the derivative of (3) with respect to α and β, evaluated at the maximum likelihood estimates. Expressions for $\underline{\widehat{\mathbf{d}}}$ and $\underline{\widehat{\mathbf{V}}}$ are presented in the Appendix.

It is important to note that the estimate of survival resulting from this over-dispersed model is larger than that under the independence model, that is, $(\alpha/(\alpha+\beta))^J$, since for each j (≥ 0) we have $(\alpha + j)/(\alpha + \beta + j) > \alpha/(\alpha + \beta)$. Thus, we would typically expect constant survival models (which assume independence of survival status at repeated visits) to underestimate overall survival in the presence of pure heterogeneity.

In some instances, the predicted probabilities of survival for individual nests $p_i^* = E[p_i \mid y_i, \ell_i, \alpha, \beta]$ may also be of interest, and these can be straightforwardly estimated by:

$$\begin{aligned}\widehat{p_i^*} &= \frac{\mathrm{B}\left(\widehat{\alpha}+\ell_i+1,\ \widehat{\beta}-y_i+1\right)}{\mathrm{B}\left(\widehat{\alpha}+\ell_i,\ \widehat{\beta}-y_i+1\right)}, \\ &= \frac{(\widehat{\alpha}+\ell_i)}{\left(\widehat{\alpha}+\ell_i+\widehat{\beta}-y_i+1\right)}.\end{aligned}$$

2.1 *Estimation of α and β*

Estimates of the shape and scale parameters of the survival probability distribution are obtained by maximizing the logarithm of the observed-data likelihood in (1) and (2). More specifically, we maximize the logarithm of:

$$\begin{aligned}L\left(\alpha,\ \beta \mid \underline{\mathbf{y}},\ \underline{\ell}\right) &= \prod_{i=1}^{n} f\left(y_i,\ \ell_i \mid \alpha,\ \beta\right), \\ &= \prod_{i=1}^{n} \int_0^1 f\left(y_i,\ \ell_i \mid p_i\right)\ f\left(p_i \mid \alpha,\ \beta\right)\ dp_i, \\ &= \prod_{i=1}^{n} \left(\frac{\mathrm{B}\left(\alpha+\ell_i,\ \beta\right)}{\mathrm{B}\left(\alpha,\ \beta\right)}\right)^{y_i} \left(\frac{\mathrm{B}\left(\alpha+\ell_i,\ \beta+1\right)}{\mathrm{B}\left(\alpha,\ \beta\right)}\right)^{1-y_i}. \qquad (4)\end{aligned}$$

where n is the number of nests in the sample, $\underline{\mathbf{y}} = \{y_1, ..., y_n\}^t$ and $\underline{\ell} = \{\ell_1, ... \ell_n\}^t$. Numerical algorithms such as Newton Rhapson may be used to maximize (4) and the asymptotic variance matrix obtained by the inverse of the negated observed information.

We now illustrate and compare our method with the constant survival model through two examples. The calculations in the examples described below were performed in the matrix language GAUSS (Aptech Systems, 1994) using programs developed by the first author.

2.2 *A Simulation Example*

We used simulation techniques to examine the behavior of the constant survival model in the presence of pure heterogeneity. A variety of biologically reasonable survival probability distributions were used. In particular, we restricted our attention to choices of α and β greater than one, so as to result in unimodal survival distributions. We believe that this may approximate more closely the pure heterogeneity that can be expected in reality. Two choices of μ_p were considered; one corresponding to a low-surviving ($\mu_p = 0.80$) and the other a high-surviving population ($\mu_p = 0.90$). Several beta distributions with mean μ_p but varying degrees of heterogeneity were considered. For each input specification, a data set of $n = 100$ nest histories was generated using the model in (1) and (2). This was repeated to generate 500 data sets. In our simulations the length of the stage (incubation or nesting) ranged from 9 till 11 days. Survival estimates for each simulated data set were calculated using the correct heterogeneous model, and the constant survival model (Mayfield, 1961, 1975; Bart and Robson, 1982):

$$L(p \mid \mathbf{y}, \ell) = \prod_{i=1}^{n} p^{\ell_i} (1-p)^{1-y_i}, \tag{5}$$

where p is the constant (across nests) probability of success. The maximum likelihood estimator of p under (5) is available in closed form and given by:

$$\hat{p} = 1 - \frac{n_f}{n_f + \sum_{i=1}^{n} \ell_i}, \tag{6}$$

where n_f is the number of failed nests. It is easy to see that (6) is simply the Mayfield (1961) estimator of survival, that is, (1 – losses/exposure), where the exposure time is calculated under the assumption that each nest failure occurred immediately prior to a visit. An estimator of the probability of surviving a period of J days, and its estimated asymptotic variance (by the delta method) under the constant model are:

$$\begin{aligned} \widehat{P_c} &= \hat{p}^J, \\ v(\widehat{P_c}) &= \hat{v}_c \left(J \, \hat{p}^{J-1} \right)^2, \end{aligned}$$

where $\widehat{v_c} = (\widehat{p}^{-2} \sum_{i=1}^{n} \ell_i + (1-\widehat{p})^{-2} n_f)^{-1}$.

Table 1 reports the results of our simulation. For each μ_p, we calculate the maximum possible variance capable of being modeled by a beta distribution. This is attained at $\beta = 1$ due to the restriction on α and β. We then choose survival distributions (that is, α and β) with 90%, 70% and 50% of this possible variation. These curves are displayed in Figure 1 for both the low and high surviving populations.

Figure 1 about here

The numbers reported in Table 1 are the average over the 500 data sets. For each of the three interval lengths, 9, 10 and 11, we present the following information: (i) average number of nests which survived (n_s); (ii) true probability of surviving the entire interval (true); (iii) average probability of survival under the heterogeneous model ($\widehat{E}\left[\widehat{\mathcal{P}}\right]$) with associated sampling standard errors, and (iv) average probability of survival under the constant model ($\widehat{E}\left[\widehat{\mathcal{P}_c}\right]$) with associated sampling standard errors. The asymptotic variance $\widehat{v}$ for estimates from both models were very comparable (always less than 0.05) and are thus not reported.

Table 1 about here

The results are fairly dramatic. As expected, constant survival models seriously underestimate overall survival at all levels of heterogeneity. For example, for the low-survival group with 90% heterogeneity, the relative bias (absolute bias/standard error) in the estimates from the heterogeneous model are 4.1, 4.7 and 5.1. The corresponding numbers for the constant survival model are 14.7, 22.2 and 29.7 respectively, which are substantially larger (almost four times that of the heterogeneous model). Problems with bias remain even at 50% heterogeneity for this group. Relative bias for estimated survival from the heterogeneous model is 1.0, 1.3 and 1.5 for the three time-units, compared with 11.4, 17 and 22 for the constant model. The effects of heterogeneity on the performance of the constant survival model are also present for the high-survival population. Relative bias in the presence of 50% heterogeneity is 0.3, 1.2 and 2 for the correct heterogeneous model and 1.6, 4.6 and 7.5 for the constant model.

The comparative magnitude of relative bias in the constant model does not diminish with increasing sample sizes. For instance, with $n = 500$ nests, we found the relative bias in the correctly specified heterogeneity model to be 1.4, 1.8 and 2 for the high-surviving population with 50% heterogeneity. For the constant model, these numbers were: 3.3, 10.5 and 17.5.

Obviously, in order to make some general statement about the bias in misspecification with a constant survival model, analytic techniques will have to be investigated. However, the results of this simulation are very conclusive in that they convey the magnitude of the problem in ignoring heterogeneity. Analysis of nest data must therefore carefully explore the possibility of inhomogeneities between nests before fitting constant survival models. In the next section we develop a likelihood ratio statistic to test for the presence of pure heterogeneity.

2.3 *Likelihood Ratio Test for Heterogeneity*

In order to formulate a hypothesis test for the presence of heterogeneity, we consider the following re-parameterization of the model in (4): $\mu_p = \alpha/(\alpha+\beta)$ and $\theta = 1/(\alpha+\beta+1)$. Then the variance of the survival probabilities p_i can be expressed as $v_p = \theta\mu_p(1-\mu_p)$. The null hypothesis of homogeneity is $H_0 : \theta = 0$ against the alternative $\theta > 0$. The likelihood function in the re-parameterized scale is:

$$L^*\left(\mu_p,\, \theta,\, |\, \underline{\mathbf{y}},\, \underline{\ell}\right) \;=\; L\left(\mu_p(1/\theta-1),\, (1-\mu)(1/\theta-1)\,|\,\underline{\mathbf{y}},\, \underline{\ell}\right),$$

and the likelihood ratio test statistic given by:

$$\Lambda \;=\; \frac{L^*\left(\widehat{\mu}_{op}, 0\,|\,\underline{\mathbf{y}},\, \underline{\ell}\right)}{L^*\left(\widehat{\mu}_p, \widehat{\theta}\,|\,\underline{\mathbf{y}},\, \underline{\ell}\right)},$$

where $\widehat{\mu}_{op}$ is the maximum likelihood estimator of μ_p under the null, $\widehat{\mu}_p$ and $\widehat{\theta}$ the estimators under the heterogeneous model. Self and Liang (1987) prove that under the null hypothesis of homogeneity, $-2\ln\Lambda$ is asymptotically distributed as a 50:50 mixture of a chi-square distribution with one degree of freedom (χ^2_1) and the constant zero. Thus, the p-value corresponding to a test

based on a χ^2_1 is halved to obtain a p-value for this one-sided test on θ. The null hypothesis of homogeneity is rejected at the 5% level of significance if the resulting p-value is smaller than 0.05.

2.4 *Hundred Acre Cove Red wings*

We now analyze nest survival data collected by Steven E. Reinert from a salt marsh population of Red-winged Blackbirds in Barrington, Rhode Island between 1982 and 1985 (Reinert, Golet and DeRagon, 1981). We focus on the nestling stage and only use the fifty three nests for which the hatching date was observed. A nest success was recorded if at least one young fledged.

Figure 2 displays the number of days from hatching till fledging for each of the 26 successful nests. The number of days for successful fledging range from 10 till 12 days.

Figure 2 about here

The trend in these data do not provide support for a constant survival assumption since the number of nest failures do not display a decreasing trend. We fit a constant survival model (equation 5) and heterogeneous model (equation 4) to these data. Estimated daily survival from the constant model is 0.93 while that from the heterogeneous model is 0.89. Maximum likelihood estimates of α and β are 5 and 0.6 respectively. Estimated survival at 12 days is 46.9% (standard error of 0.068) for the heterogeneous model and 43.0% (standard error of 0.069) for the constant survival model. Figure 3 displays the probability of survival for varying interval lengths using both models.

Figure 3 about here

The survival curve from the heterogeneous model is much flatter, especially for the larger interval lengths, compared to the negative exponential curve of the constant survival model. The heterogeneous model yields a log-likelihood of -97.496 with one additional parameter, compared with -98.709 from the constant survival model. The likelihood ratio statistic to test the null hypothesis of homogeneity is 2.426 which corresponds to a p-value of 0.050, as explained in Section 2.3. Thus, there appears to be evidence of heterogeneity among the nests. It is conceivable that some of

this heterogeneity may be explained by adjusting for potentially important covariates; Section 3.2 investigates exactly such an analysis.

A primary advantage of the beta framework described in Section 2 is that variability between nests is modeled directly through the success probabilities p_i. However, it suffers from the limitation that it does not explain sources of heterogeneity. In Section 3 we formulate a logistic normal model to estimate nest survival in this setting.

3 Logistic Normal Model for Nest Survival

In this section we describe a very general class of random-effect models to estimate nest success in the presence of one or more sources of heterogeneity. Conditional on the survival probabilities p_i the contribution of nest i to the observed likelihood is the same as in (1). However, the nest-specific probabilities p_i are modeled as:

$$\text{logit}\{p_i\} = \underline{\mathbf{x}}_i^t \underline{\gamma} + v_i, \tag{7}$$

where $\underline{\mathbf{x}}_i$ is a $p \times 1$ vector of covariates for nest i, $\underline{\gamma}$ is a $p \times 1$ vector of unknown parameters and v_i is a nest-specific intercept. (Note that a slight change in the formulation of (7) can incorporate time-varying covariates as described in equation (11).) The parameter $\underline{\gamma}$ captures the effect of the covariates on survival for the average nest in the population, while v_i modifies the average response to make it specific to nest i. The specification in (7) is completed by postulating:

$$v_i = \underline{\mathbf{z}}_i^t \underline{\mathbf{b}} + u_i, \tag{8}$$

where $\underline{\mathbf{z}}_i$ is a $q \times 1$ design vector for the random-effects $\underline{\mathbf{b}}$ and u_i is random error. The random-effects $\underline{\mathbf{b}}$ can denote variations due to site, geographical location, year of data collection or other tangible sources of heterogeneity. (It is prudent to entertain such models only when there are a reasonable number of levels of the heterogeneity source. Opinions vary, but generally five or more levels are considered sufficient to investigate such models.) Typical distributional assumptions are $\underline{\mathbf{b}} \sim \mathcal{N}_q(0, \underline{\mathbf{D}})$ independentally of $u_i \sim \mathcal{N}(0, \sigma^2)$. Thus, equation (8) decomposes the variation

between nests into pure heterogeneity (that is, σ^2) and that explained by differences in identifiable factors such as site (that is, $\underline{\mathbf{D}}$). This decomposition induces a correlation between survival status for a nest on repeated occasions as before, but in addition, it also induces correlation between nests which share a random-effect. Models such as these have been studied extensively for analyzing spatially correlated data and are encompassed within the realm of spatial hierarchical models (see Ghosh, et al., 1997).

Often, in practice, ornithologists may be interested in estimating survival in the presence of a single explainable source of heterogeneity, say k plots. Then equation (8) resembles the formulation for an analysis of variance, namely,

$$v_{ij} \;=\; b_i + u_{ij}, \qquad i \;=\; 1,\ldots,k,\;\; j \;=\; 1,\ldots,r_i, \tag{9}$$

where i indexes plots, j indexes nest within plots, b_i is a plot-specific random-intercept distributed as $b_i \sim \mathcal{N}(0,\, \theta)$ and $u_{ij} \sim \mathcal{N}(0,\, \sigma^2)$. Equation (9) arises from equation (8) by defining $\underline{\mathbf{Z}} = \underline{\mathbf{I}} \otimes \underline{\mathbf{1}}$ where $\underline{\mathbf{Z}}$ is the $n \times q$ matrix with rows $\underline{\mathbf{z}}_i$, $\underline{\mathbf{I}}$ the identity matrix, $\underline{\mathbf{1}}$ a vector of ones and $\otimes$ the direct product operator. A variety of other heterogeneity patterns may be modeled by appropriate choices of $\underline{\mathbf{Z}}$.

Conditional on the random-effects $\underline{\mathbf{b}}$, nests are statistically independent with likelihood given by (1). Unconditionally, the likelihood is given by:

$$L(\underline{\gamma}, \sigma^2, \underline{\mathbf{D}} \mid \underline{\mathbf{y}},\, \underline{\ell}) \;\propto\; \int \left\{ \prod_{i=1}^{n} \int f\,(y_i, \ell_i \mid u_i,\, \underline{\mathbf{b}})\, \frac{exp\left(-\frac{u_i^2}{2\sigma^2}\right)}{\sigma}\, du_i \right\} \frac{exp\left(-\frac{1}{2}\underline{\mathbf{b}}^t \underline{\mathbf{D}}^{-1} \underline{\mathbf{b}}\right)}{|\underline{\mathbf{D}}|^{1/2}}\, d\underline{\mathbf{b}}. \tag{10}$$

Closed form expressions do not exist for the likelihood in (10). However, for simple random-effect structures (that is, $\underline{\mathbf{Z}}$), numerical integration methods may be used to provide an estimate. (Hedeker and Gibbons (1994) have developed several programs to estimate integrals of the form in (10)) For complicated models the Monte Carlo method may be used to provide a simulation-based estimate of the likelihood function. (McCulloch (1997) has studied the quality of various Monte Carlo approximants for related problems.) This estimated likelihood can then be maximized to obtain

maximum likelihood estimates of the regression parameters $\underline{\gamma}$, and the components of variance σ^2 and $\underline{\mathbf{D}}$.

As in Section 2 the inferential goal lies in estimating the unconditional (integrated over the random-effects) probability of surviving J days, which in this context is calculated for specific covariate strata $\underline{\mathbf{x}}$, that is,

$$\mathcal{P}(\underline{\mathbf{x}}) = \frac{1}{(2\pi)^{\frac{q+1}{2}}} \int \left\{ \int p_i^J \frac{exp\left(-\frac{u_i^2}{2\sigma^2}\right)}{\sigma} du_i \right\} \frac{exp\left(-\frac{1}{2}\underline{\mathbf{b}}^t \underline{\mathbf{D}}^{-1} \underline{\mathbf{b}}\right)}{|\underline{\mathbf{D}}|^{1/2}} d\underline{\mathbf{b}},$$

where p_i is given by equations (7) and (8). The above expression involves calculations similar to those required for likelihood evaluation. Asymptotic standard errors may be obtained by the delta method.

Often, interest may also focus on predicted probabilities of survival for individual nests, which is given by $E\left[\left(1 + exp\left(-\underline{\mathbf{x}}_i^t \underline{\gamma} - \underline{\mathbf{z}}_i^t \underline{\mathbf{b}} - u_i\right)\right)^{-J} \mid \underline{\mathbf{y}}, \underline{\ell}\right]$ with the expectation over the conditional distribution $f\left(\underline{\mathbf{b}}, \underline{\ell} \mid \underline{\mathbf{y}}, \underline{\ell}\right)$. Predicted values of individual random-effects may also be calculated as $\widehat{\underline{\mathbf{b}}} = E\left[\underline{\mathbf{b}} \mid \underline{\mathbf{y}}, \underline{\ell}\right]$ and $\widehat{u}_i = E\left[u_i \mid \underline{\mathbf{y}}, \underline{\ell}\right]$.

In the next section we will describe some of the calculations for a single source of explainable heterogeneity, that is, for $\underline{\mathbf{Z}} = \underline{\mathbf{I}} \otimes \underline{\mathbf{1}}$.

3.1 *Single source of explainable heterogeneity*

For a design with k levels of a source of heterogeneity and r_i nests per level, the likelihood in (10) reduces to:

$$L(\underline{\gamma}, \theta, \sigma^2 \mid \underline{\mathbf{y}}, \underline{\ell}) \propto \prod_{i=1}^{k} \int \left\{ \prod_{j=1}^{r_i} \int f\left(y_{ij}, \ell_{ij} \mid u_{ij}, b_i\right) \frac{exp\left(-\frac{u_{ij}^2}{2\sigma^2}\right)}{\sigma} du_{ij} \right\} \frac{exp\left(-\frac{1}{2\theta} b_i^2\right)}{\sqrt{\theta}} db_i.$$

An estimate of the above likelihood may be obtained very accurately using quadrature methods (Abramowitz and Stegun, 1964), even for k as large as 100, and is given by the expression $exp\left(\sum_{i=1}^{k} \ln\left\{\sum_h w_h \, exp\left[\sum_{j=1}^{r_i} \ln\{\sum_g w_g \, p_{hg}^{\ell_{ij}} \, (1 - p_{hg})^{1-y_{ij}}\right\}\right]\right)$, where the probability $p_{hg} = \left(1 + exp\left(-\underline{\mathbf{x}}_{ij}^t \underline{\gamma} - \sqrt{\theta} a_h - \sqrt{\sigma^2} a_g\right)\right)^{-1}$, and a, w are the abscissae and weights associated with

Gauss-Hermite quadrature. (Note that we calculate logarithms and then exponentiate the resulting expressions, in order to avoid underflow errors.) Overall unconditional survival for a particular covariate strata $\underline{\mathbf{x}}$ is:

$$\mathcal{P}(\underline{\mathbf{x}}) \approx \frac{1}{(2\pi)^{\frac{k+1}{2}}} \sum_h w_h \sum_g w_g \, p_{hg}^J.$$

The null hypothesis of homogeneity for the oneway random-effects model is $H_0 : \theta = 0$, versus the alternative $\theta > 0$. As in Section 2.3 a likelihood ratio statistic to perform this test is given by:

$$\Lambda = \frac{L\left(\underline{\widehat{\gamma}}_o, \widehat{\sigma_o^2}, 0 \mid \underline{\mathbf{y}}, \underline{\ell}\right)}{L\left(\underline{\widehat{\gamma}}, \widehat{\sigma^2}, \widehat{\theta} \mid \underline{\mathbf{y}}, \underline{\ell}\right)},$$

where $\underline{\widehat{\gamma}}_o$, $\widehat{\sigma_o^2}$ are the maximum likelihood estimates under the null, and $\underline{\widehat{\gamma}}$, $\widehat{\sigma^2}$ and $\widehat{\theta}$ are the unrestricted estimates. The null hypothesis of homogeneity is rejected for large positive values of $-2 \ln \Lambda$, as compared to a 50:50 mixture of a χ_1^2 and the constant zero.

3.2 *Red-winged Blackbirds Revisited*

In order to illustrate the use of logit-normal models, we explore an analysis of the Red-wing data to (i) adjust for potential effects of age on daily survival, and (ii) account for suspected variations across the four data collection years. More specifically, in year i ($i = 1, \ldots, 4$), for nest j at age t, we postulate the following model:

$$\text{logit}\,\{p_{ijt}\} = \gamma_0 + \gamma_1 \, I(t \leq 2) + \gamma_2 \, I(3 \leq t \leq 6) + \gamma_3 \, I(t \geq 7) + b_i + u_{ij}, \qquad (11)$$

where $I(.)$ is the indicator function, b_i a year-specific random-intercept and u_{ij} is a nest-within-year specific intercept. The above formulation allows daily survival to vary as a function of nest age which is grouped into three intervals. (We explored other choices for the age intervals, including finer and coarser divisions, but did not notice any substantial differences in our results.) Table 2 compares estimated probability of 12-day survival rates and their standard errors for five models: constant survival model, Pollock-Cornelius full model (using twelve age-specific failure probabilities, one for each day in the nestling period), an independence logit model with age-effects (that is, equation (11)

without the random-effects b_i and u_{ij}), a logit-normal model with age-effects and nest intercepts alone (that is, equation (11) without the year-specific intercepts b_i) and the full logit-normal model in equation (11). By incrementally adding the nest and year intercepts we hope to isolate their individual contributions and develop a parsimonious model to best explain the variations in these data. We used the estimated likelihood and survival estimates to guide us in our choice of a model.

Estimated 12-day survival from the constant model is 43% and 45% from the independence logit model compared with 50% from both the random-effect models and the Pollock-Cornelius model. It is important to note that the survival estimates from the random-effect models are interpreted as average survival for an entire population of Red-wing Blackbirds, while those from the constant, independence logit and Pollock Cornelius models pertain to the specific nests under observation. Addition of nest-specific intercepts to the independence logit model (Model III in Table 2) results in a significant increase in the likelihood ($p = 0.031$). This suggests that discrepancies in survival propensity remain over and above age influences, which, if not accounted for, can lead to models that underestimate overall survival. Inclusion of year-specific intercepts to the model adjusted for age and nest-level variation (Model IV in Table 2) does not provide a significant improvement in fit ($p = 0.16$), suggesting the absence of correlation in survival within years. (This may be a manifestation of the small number of years under study.) It is thus reasonable to pool the data over years and consider a model with age and nest effects alone. A goodness of fit test of this reduced model with the saturated Pollock-Cornelius model shows that the latter (despite having 8 additional parameters and making no assumptions about the functional form of survival probabilities) does not provide a significantly better fit ($p=0.99$). Thus, by carefully accounting for inhomogeneities between nests, we have developed a parsimonious model which can capture the effect of covariates as well as correlations induced at various levels. Average daily survival estimates (and standard errors) for each of the three age-intervals from the logit-normal model with age and nest effects are: 0.880 (0.035), 0.772 (0.073) and 0.724 (0.067).

4 Extensions to Encounter Sampling

Typically, in practice, nests are found at various stages of their development. Thus, the observed data is the pair $\{y_i,\ t_i\}$, where y_i is as defined in Section 2, and t_i is the number of time-units a nest is under observation and active. This observed interval is often smaller than the true length of life ℓ_i. However, assuming that the time required for a nest to succeed is a known constant $\mathcal{L}$ these observed data contain information on ℓ_i. Thus, for instance, $\ell_i = \mathcal{L}$ for nests that succeed and $t_i \leq \ell_i \leq (\mathcal{L} - 1)$ for nests that fail. Ignoring this additional information and computing survival estimates based only on the observation time can result in conservative estimates of nesting success. Modeling extensions to encompass such encounter sampling is straightforward both for the pure heterogeneity model and the logistic-normal models.

4.1 *Pure heterogeneity model for encounter sampling*

Conditional on the nest-specific probabilities p_i, the contribution of nest i to the observed likelihood is:

$$f\left(y_i,\ t_i \mid p_i\right) \quad = \quad \begin{cases} f\left(t_i \mid \ell_i = \mathcal{L},\ \varphi\right) p_i^{\mathcal{L}} & \text{if} \quad y_i = 1, \\ \sum_{\ell=t_i}^{\mathcal{L}-1} f\left(t_i \mid \ell_i = \ell,\ \varphi\right) p_i^{\ell}\left(1 - p_i\right) & \text{if} \quad y_i = 0. \end{cases} \tag{12}$$

The distribution $f\left(t_i \mid \ell_i,\ \varphi\right)$ in equation (12) is determined by the search strategy used to sample nests. The sampling schemes commonly used in practice are discussed by Bromaghin and McDonald (1993); for illustration purpose, we consider the systematic sampling scheme, that is,

$$f\left(t_i \mid \ell_i,\ \varphi\right) \quad = \quad \frac{\varphi(1-\varphi)^{\ell_i - t_i}}{\left(1 - (1-\varphi)^{\ell_i}\right)},$$

where φ is the daily probability of detection, which is assumed constant across nests. Bromaghin and McDonald (1993) also discuss the issue of weighting the joint distribution of survival and lifetimes $f\left(y_i,\ \ell_i\right)$ to account for the probability sampling inherent in nest data: these ideas can be applied in our context as well. It is easy to see that the unconditional observed-data likelihood is $\left(\prod_{i=1}^{n}\left(f\left(t_i \mid \ell_i = \mathcal{L},\ \varphi\right) \mathrm{B}\left(\alpha + \mathcal{L},\ \beta\right)\right)^{y_i} \left(\sum_{\ell=t_i}^{\mathcal{L}-1} f\left(t_i \mid \ell_i = \ell,\ \varphi\right) \mathrm{B}\left(\alpha + \ell,\ \beta + 1\right)\right)^{1-y_i}\right) / \mathrm{B}\left(\alpha,\ \beta\right)$.

Estimation in this setting can proceed as before, by numerically maximizing the likelihood function to obtain maximum likelihood estimates of the regression parameters, components of variance and detection parameters.

4.2 *Logistic-normal model for encounter sampling*

Conditional on the random-effects $\underline{\mathbf{b}}$ and pure heterogeneity u_i, the contribution of nest i to the observed likelihood is as given in equation (12) with p_i of the form given by (7) and (8). The unconditional observed-data likelihood $L\left(\underline{\gamma},\ \varphi,\ \underline{\mathbf{D}},\ \sigma^2 \mid \underline{\mathbf{y}},\ \underline{\mathbf{t}}\right)$ is proportional to

$$\int \left\{ \prod_{i=1}^{n} \int f\left(y_i,\ t_i \mid \underline{\mathbf{b}},\ u_i\right) \frac{exp\left(-\frac{1}{2\sigma^2}u_i^2\right)}{\sigma}\, du_i \right\} \frac{exp\left(-\frac{1}{2}\underline{\mathbf{b}}^t \underline{\mathbf{D}}^{-1} \underline{\mathbf{b}}\right)}{|\underline{\mathbf{D}}|^{1/2}}\, d\underline{\mathbf{b}},$$

and can be estimated in a similar fashion as the logistic-normal models.

5 Conclusion

This paper presents a flexible random-effects framework to accommodate heterogeneities in nest survival data. We recommend that analysis of nest data begin with an examination of models which account for pure heterogeneity. If data has been collected from several different locations (or multiple years), models which induce correlation between nests from the same location (or year) should also be considered. If there appear to be insufficient evidence of disparities based on these analyses, only then should one resort to constant survival models.

Acknowledgment

We thank an Associate Editor and anonymous referee for several comments which greatly improved the manuscript. We also thank Steven E. Reinert, M.S., Medical Computing, LifeSpan, Rhode Island, for providing the data analyzed here. The work of the second author was supported by NSF grant DMS-9625476.

References

Abramowitz, M. and Stegun, I. (1964). *Handbook of Mathematical Functions.* U.S. Government Printing Office, Washington DC.

Bart, J. and Robson, D. S. (1982). Estimating survivorship when the subjects are visited periodically. *Ecology* **63,** 1078-1090.

Bromaghin, J. F. and McDonald, L. L. (1993). Weighted nest survival models. *Biometrics* **49,** 1164-1172.

Burnham, K. P., and Rexstad, E. A. (1993). Modeling heterogeneity in survival rates of banded waterfowl. *Biometrics* **49,** 1194-1208.

GAUSS, Aptech Systems, Inc. Maple Valley WA 1984-1994.

Ghosh, M., Natarajan, K., Stroud, T. W. F. and Carlin, B. P. (1997). Generalized linear models for small area estimation. *Journal of the American Statistical Association*, In press.

Griffiths, D. A. (1973). Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics* **29,** 637-648.

Hedeker, D. and Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics* **50,** 933-944.

Johnson, D. H. (1979). Estimating nest success: The Mayfield method and an alternative. *The Auk* **96,** 651-661.

Kahn, M. J. and Raftery, A. E. (1996). Discharge rates of Medicare stroke patients to skilled nursing facilities: Bayesian logistic regression with unobserved heterogeneity. *Journal of the American Statistical Association* **91,** 29-41.

Klett, A. T. and Johnson, D. H. (1982). Variability in nest survival rates and implications to nesting studies. *The Auk* **99,** 77-87.

Mayfield, H. (1961). Nesting success calculated from exposure. *Wilson Bulletin* **73,** 255-261.

Mayfield, H. (1975). Suggestions for calculating nest success. *Wilson Bulletin* **87,** 456-466.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* **92,** 162-170.

Pollock, K. H. and Cornelius, W. L. (1988). A distribution-free nest survival model. *Biometrics* **44,** 397-404.

Reinert, S. E., Golet, F. C. and DeRagon, W. R. (1981). Avian use of ditched and unditched salt marshes in southeastern New England: a preliminary report. *Proceedings Northeast Mosquito Control Association* **27,** 1-23.

Self, S. G., and Liang, K-Y (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82,** 605-610.

Williams, D. A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* **31,** 949-952.

Appendix

Expression for $\mathbf{d}(\alpha, \beta)$

$$\mathbf{d}(\alpha, \beta) = \mathcal{P}\begin{pmatrix} \beta \sum_{j=1}^{J} ((\alpha + J - j)\ (\alpha + \beta + J - j))^{-1} \\ -\sum_{j=1}^{J} (\alpha + \beta + J - j)^{-1} \end{pmatrix}$$

where $\mathcal{P}$ is defined in equation (3).

Expression for $V(\alpha, \beta)$

The asymptotic variance matrix $V(\alpha, \beta) = \mathrm{I}^{-1}(\alpha, \beta)$ where $I(.)$ is the negated observed information:

$$I(\alpha, \beta) = -\begin{pmatrix} i_{11} & i_{12} \\ i_{12} & i_{22} \end{pmatrix}$$

where

$$i_{11} = \sum_{i=1}^{n} \left[y_i \left\{ -\sum_{j=1}^{\ell_i} \frac{1}{(\alpha + \ell_i - j)^2} + \sum_{j=1}^{\ell_i} \frac{1}{(\alpha + \beta + \ell_i - j)^2} \right\} + (1 - y_i) \left\{ -\sum_{j=1}^{\ell_i} \frac{1}{(\alpha + \ell_i - j)^2} + \sum_{j=0}^{\ell_i} \frac{1}{(\alpha + \beta + \ell_i - j)^2} \right\} \right],$$

$$i_{12} = \sum_{i=1}^{n} \left[y_i \left\{ \sum_{j=1}^{\ell_i} \frac{1}{(\alpha + \beta + \ell_i - j)^2} \right\} + (1 - y_i) \left\{ \sum_{j=0}^{\ell_i} \frac{1}{(\alpha + \beta + \ell_i - j)^2} \right\} \right],$$

$$i_{22} = \sum_{i=1}^{n} \left[y_i \left\{ \sum_{j=1}^{\ell_i} \frac{1}{(\alpha + \beta + \ell_i - j)^2} \right\} + (1 - y_i) \left\{ -\frac{1}{\beta^2} + \sum_{j=0}^{\ell_i} \frac{1}{(\alpha + \beta + \ell_i - j)^2} \right\} \right].$$

Captions for Tables

Caption for Table 1

A comparison of average survival calculated from the constant survival model $\widehat{E}\left[\widehat{\mathcal{P}_c}\right]$ and a pure heterogeneity model $\widehat{E}\left[\widehat{\mathcal{P}}\right]$ for various survival distributions. The estimates reported are the average over 500 data sets, each with 100 nests. Sampling standard errors across the 500 data sets are reported in parentheses. For each specification, the true probabilities of survival (true), average number of nests which survived n_s, mean μ_p and variance v_p of the survival distributions are displayed. The variance v_p is also expressed as a percentage of the maximum variance (labeled % heterogeneity) allowed under a beta distribution with mean μ_p. Sampling standard errors for n_s are always less than 0.17.

Caption for Table 2

Comparison of estimated twelve-day survival and their standard errors for the Red-winged Blackbird data for five models.

Captions for Figures

Caption for Figure 1

Beta survival distributions with varying levels (50%, 70%, 90%) of heterogeneity for (a) a low surviving (mean daily survival $\mu_p = 0.80$) and (b) a high surviving (mean daily survival $\mu_p = 0.90$) population.

Caption for Figure 2

Distribution of number of days (from hatching) for successful fledging of Red-winged Blackbird nests.

Caption for Figure 3

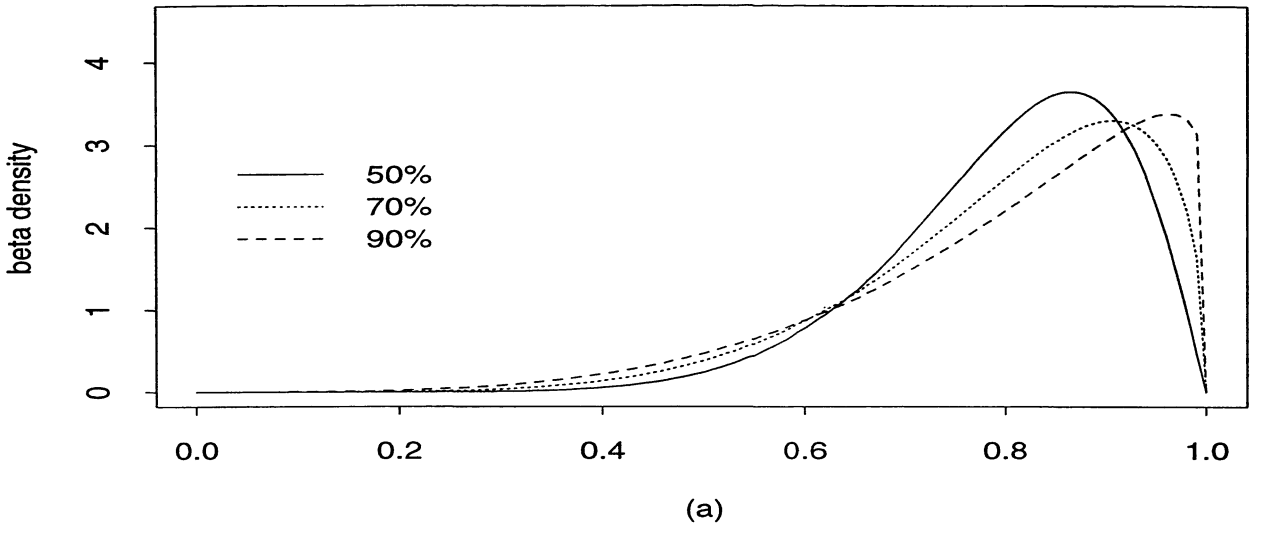
Estimated probability of nest survival (nestling stage) for Red-winged Blackbird nests using the heterogeneous and constant survival model.

Table 1

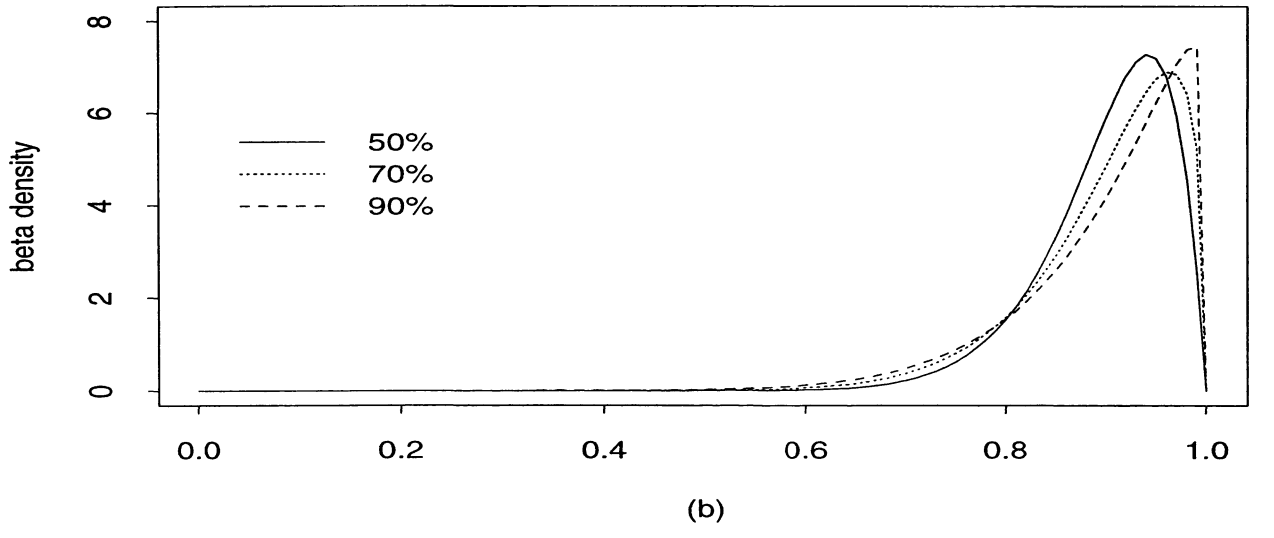
v_p	% heterogeneity	*Interval Length*	n_s	*True*	$\widehat{E}\left[\widehat{\mathcal{P}}\right]$		$\widehat{E}\left[\widehat{\mathcal{P}_c}\right]$	
Case 1: Low surviving population $\left(\mu_p = 0.80,\ \max v_p = \frac{\mu_p(1-\mu_p)^2}{2-\mu_p} = 0.027\right)$								
		9	9.674	0.292	0.285	(0.002)	0.263	(0.002)
0.024	90%	10	8.790	0.270	0.262	(0.002)	0.227	(0.002)
		11	8.550	0.250	0.241	(0.002)	0.196	(0.002)
		9	8.390	0.261	0.260	(0.002)	0.236	(0.002)
0.019	70%	10	7.968	0.237	0.235	(0.002)	0.202	(0.002)
		11	7.514	0.217	0.215	(0.002)	0.172	(0.002)
		9	7.714	0.227	0.229	(0.002)	0.207	(0.002)
0.013	50%	10	6.678	0.202	0.205	(0.002)	0.174	(0.002)
		11	6.266	0.181	0.184	(0.002)	0.147	(0.001)
Case 2: High surviving population $\left(\mu_p = 0.90,\ \max v_p = \frac{\mu_p(1-\mu_p)^2}{2-\mu_p} = 0.008\right)$								
		9	16.362	0.491	0.488	(0.002)	0.487	(0.002)
0.007	90%	10	15.662	0.463	0.460	(0.002)	0.449	(0.002)
		11	14.732	0.439	0.434	(0.002)	0.415	(0.002)
		9	15.558	0.471	0.472	(0.002)	0.469	(0.002)
0.006	70%	10	14.848	0.442	0.444	(0.002)	0.432	(0.002)
		11	14.520	0.416	0.418	(0.002)	0.398	(0.002)
		9	14.976	0.450	0.450	(0.002)	0.446	(0.002)
0.004	50%	10	14.210	0.418	0.421	(0.002)	0.408	(0.002)
		11	13.316	0.390	0.394	(0.002)	0.374	(0.002)

Table 2

Model	No. parameters	Survival	Standard error	Log likelihood
I. Constant survival model	1	0.430	0.069	-98.709
II. Logit model with age-effects	3	0.456	0.072	-97.922
III. Logit-normal model with age and nest effects	4	0.504	0.082	-96.225
IV. Logit-normal model with age, nest and year effects	5	0.490	0.095	-95.760
V. Pollock-Cornelius model	12	0.490	0.070	-95.798


beta density
0
1
2
3
4
50%
70%
90%
0.0
0.2
0.4
0.6
0.8
1.0

(a)


beta density
0
2
4
6
8
50%
70%
90%
0.0
0.2
0.4
0.6
0.8
1.0

(b)

Figure 1

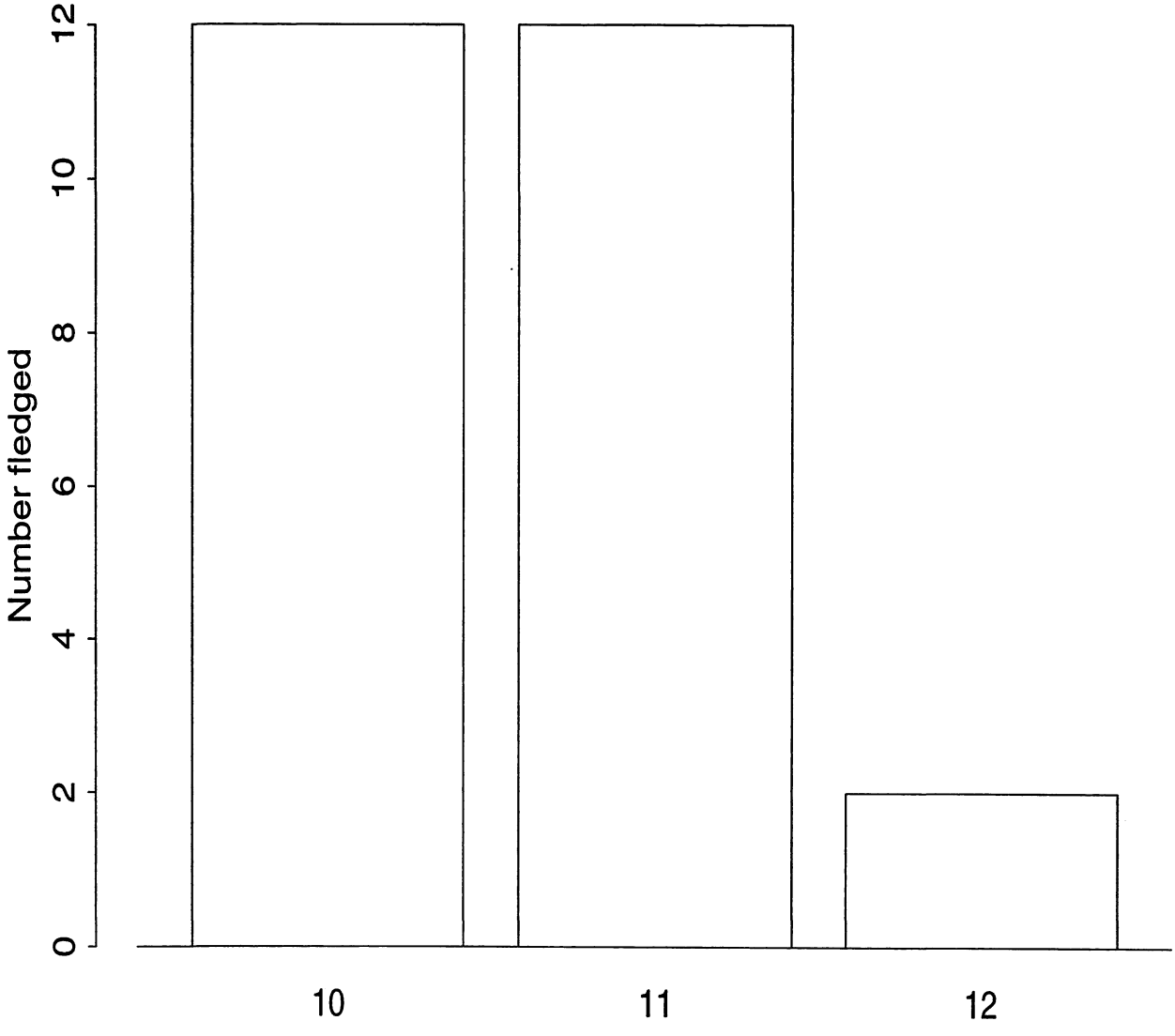


Figure 2

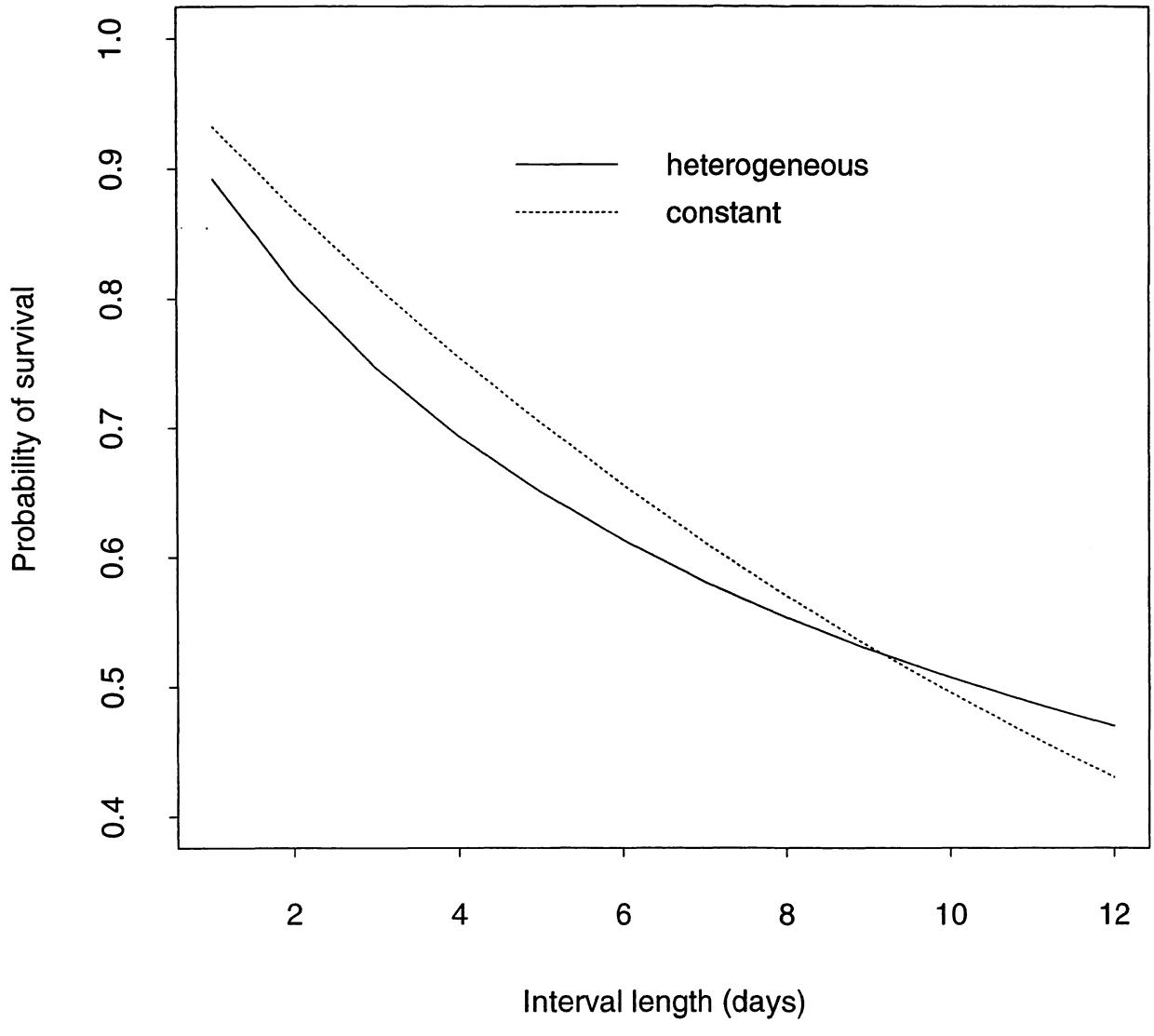

heterogeneous
constant
Probability of survival
1.0
0.9
0.8
0.7
0.6
0.5
0.4
2
4
6
8
10
12
Interval length (days)

Figure 3